Learning Activity Features of High Performance Students

Fumiya Okubo, Kyushu University, Japan, fokubo@artsci.kyushu-u.ac.jp
Sachio Hirokawa, Kyushu University, Japan, hirokawa@cc.kyushu-u.ac.jp
Misato Oi, Kyushu University, Japan, oimisato@gmail.com
Atsushi Shimada, Kyushu University, Japan, atsushi@limu.ait.kyushu-u.ac.jp
Kentaro Kojima, Kyushu University, Japan, kojima@artsci.kyushu-u.ac.jp
Masanori Yamada, Kyushu University, Japan, mark@mark-lab.net
Hiroaki Ogata, Kyushu University, Japan, hiroaki.ogata@gmail.com

Abstract: In this paper, we present a method of identifying learning activities that are important for students to achieve good grades. For this purpose, the data of 99 students were collected from a learning management system and an e-book system, including attendance, time on preparation and review, submission of reports, and quiz scores. We applied a support vector machine to these data to calculate a score of importance for each learning activity reflecting its contribution to the attainment of an A grade. Selecting certain important learning activities by following several evaluation measures, we verified that these learning activities played a crucial role in predicting final student achievements. One of the obtained results implies that time on preparation and review in the middle part of a course influences a student's final achievement.

Keywords: learning analytics, data of LMS and e-book system, learning activity feature, support vector machine

Introduction

In recent years, intensive data mining in the field of education research has become possible, as the widespread use of ICT-based educational systems, such as learning management systems (LMSs) and e-book systems. These systems enable us to automatically collect many kinds of log data that corresponding to students' online learning activities. Such collected data can be analyzed in order to identify the students' learning activities and typical learning patterns of particular target students or groups, for example, those who are likely to fail or drop out of class, commonly referred to as "at-risk" students (Baradwaj & Pal (2011)).

In October 2014, the well-known LMS Moodle and the e-book system BookLooper(1), provided by KYOCERA MARUZEN Systems Integration Co., Ltd., were introduced at Kyushu University in Japan in order to facilitate the collection and analysis of educational data. The e-book system records detailed action logs with the user id and timestamp, such as moves back and forth between pages, the contents of memos, and the kind of access device used (PC or smartphone). These records enable us to investigate a range of learning activities, both inside and outside of class. Utilizing the log data stored in Moodle and BookLooper, a number of investigations have been conducted at Kyushu University (Ogata et al. (2015)).

Several studies on educational data mining have some specified measures of learning activities that are utilized for visualizing and analyzing students' learning behaviors. For example, in Okubo et al. (2015), it has realized to visualize the four types of learning logs stored in an LMS and an e-book system, namely attendance, time spent for browsing slides in an e-book system, submission of reports and quiz scores, by using a discrete graph for each academic achievement, referring to the method proposed in Hlosta et al. (2014). On the other hand, You (2016) has claimed that researchers ought to identify meaningful learning activities to predict students' achievement. They have verified significance of particular learning activities by the statistical analysis methods. Focusing on methods of analysis, Ifenthaler and Widanapathirana (2014) showed case studies of educational data mining utilizing the well-known method of machine learning, namely support vector machines (SVMs) introduced in Cortes & Vapnik (1995). Goda et al. (2013) applied an SVM to students' comments and showed the relationships between self-evaluation comments and the final grade of the students.

In this paper, we propose a method of discovering learning activities that are important for students to achieve good grades, by applying SVM to log data stored in an LMS and an e-book system. For this purpose, we use four types of learning logs, namely, attendance, slide browsing time, submission of reports and quiz scores. We verify the performance of prediction of student's final achievement on the basis of only certain learning activities selected as important by our method. We also discuss an interpretation on learning activities selected as important by our method. Finally, we give an indication of future research plans along the same line as the research presented in this paper.

Method

Data collection

We collected the learning logs of 99 students attending an "Information Science" course that started in October, 2014. The course was held over 14 weeks, with a cancellation in the 9th week: thus, 13 lectures were given. Each of these lectures was presented by using several slides in the e-book system, with each slide associated with only one lecture. The slides were used by the students to complete their preparation and/or review sessions before and after each lecture, respectively. Furthermore, the students were required each week to submit a report and answer a quiz that contained three to five questions related to that week's lecture through the LMS.

As mentioned above, we refer to four kinds of data stored in the LMS and the e-book system in this paper, namely

- attendance or absence (represented by p),
- submission of a report or failure to do so (r),
- a sum of the time spent browsing slides for preparation and/or review (b), and
- quiz score (t),

of each student participating in each week of the course. For each of the four items, we consider whether or not it was achieved. Thus, for the ith week, a particular student's lecture attendance or absence is coded by the word *ip:o* or *ip:x*, respectively. Each student's report submission datum was coded as *ir:o* if a report was submitted and *ir:x* if not, respectively. Slide browsing time was also transformed into a binary category, with browsing time of 600 seconds or longer coded as *ib:o*, and anything shorter as *ib:x*. Similarly, a quiz score of 70% or above was coded as *it:o*, and anything lower as *it:x*. For example, if a student in the 7th week attended a lecture, did not submit a report, browsed slides for longer than 600 seconds, and scored below 70% on the quiz, the words *7p:o*, *7r:x*, *7b:o*, and *7t :x* would represent this student's activities in the 7th week. We note that
the data on the activities of each student in the class can be represented by an 8*14=112-dimensional vector, in which each element is either 0 (for *ip:x, ir:x, ib:x,* and *it:x*) or 1 (for *ip:o, ir:o, ib:o,* and *it:o*).

The students in the course were graded in terms of the categories A, B, C, D, and F according to their grade score out of 100, which reflected all their activities. The relationship between grade and grade score is indicated in Table 1, which shows the frequency with which the 99 students attained each grade. The words *s:A, s:B, s:C. s:D* and *s:F* were used to represent the log data of these grades.

By registering the data of the 99 students as documents, we constructed a search engine by means of GETA[(2)](Generic Engine for Transposable Association) provided by National Institute of Informatics.

Table 1: Frequency of grades and grade scores.

Grade	Grade score range	Number of students
A	90-100	37
B	80-89	30
C	70-79	13
D	60-69	9
F	0-59	10

Classification based on SVM and feature selection

The aim of this study is to discover important learning activities that distinguish students achieving A grades from students achieving lower grades. For this purpose, we utilized an SVM in which the documents of A grade students are positive instances and the documents of students with other grades are negative instances. Following the method proposed in Sakai & Hirokawa (2012), we applied machine learning method based on SVM and feature selection to classify students' learning activities data.

An evaluation of classification performance of the proposed method is conducted be means of 5-fold cross validation. Thus, the data are separated into five parts, of which four parts are used as training data and the remaining part as test data. Then, there are five ways to choose the parts for training data and test data. Thus, the final result of such 5-fold cross validation is the average of the results of these five ways.

Specifically, our method is conducted in terms of the following steps.

Data generation by multiple instance method

The collected data consisted of 37 positive instances and 62 negative instances. As the number of instances in the training data is somewhat small, we attempt to apply a restricted version of multiple instance learning

(Dietterich et al. (1997)). Specifically, from the training data, we randomly choose a pair of positive instances, and regard this pair as a new positive instance (called a “bag” in Dietterich et al. (1997)). In this way, we construct new 100 positive instances. Similarly, 100 new negative instances can be constructed. By adding these new instances to the original training data, we can increase its volume.

Application of linear SVM

Applying linear SVM to the training data containing all words, we construct the model that classifies the documents of A grade students. We used SVM-light[3] for the learning tool.

Recall that a document of a student is represented by a 112-dimensional vector. For a word *w* and a document *d*, the number of occurrences of *w* in *d* is denoted by *tf(w, d),* which equals either 0 or 1. The classifier (or model) *f* of the linear SVM learned from the training data is of the form

$$f(d) = \underset{w_i \Box d}{\Box}\ weight(w_i)\Box tf(w_i, d) + b,$$

where *b* is a constant term. For a document *d* of test data, if *f(d)* is greater than 0, then *d* is classified as a document of grade A student. Conversely, if *f(d)* is less than 0, then *d* is classified as a document of a student with another grade.

Score of word and feature selection

For a given word w_i, its *weight*(w_i) can be regarded as a score of importance of w_i on the classifier *f.* A score of importance of w_i can be obtained by applying *f* to a document containing only w_i and removing the constant *b*.

Feature selection of words is conducted by following the six measures for evaluation, shown in Table 2, in which the number of documents containing *w* is denoted by *df(w)* and an absolute value of *x* is denoted by *abs(x)*. For *N=1, 2, ..., 10, 20, ...70*, we choose the top *N* positive words and top *N* negative words regarding each measure of *w.o, d.o,* and *l.o*. Similarly, we choose the top *2N* words regarding each measure of *w.a, d.a,* and *l.a*. These *2N* words are called feature words of *f.*

We apply linear SVM to the training data which containing *2N* words selected by the six type of feature selection and to the training data of all words. Then, using the test data, we evaluate an estimation performance of each model obtained by the linear SVM, by means of 5-fold cross validation.

Table 2: Measures for feature selection.

Type	Measure for evaluation
w.o	*weight*(w_i)
d.o	*weight*(w_i)**df*(w_i)
l.o	*weight*(w_i)**log*(*df*(w_i))
w.a	*abs*(*weight*(w_i))
d.a	*abs*(*weight*(w_i)**df*(w_i))
l.a	*abs*(*weight*(w_i)**log*(*df*(w_i)))

Experimental results

Accuracy

We have conducted experiments to obtain the values for accuracy, precision, recall, and F-measure as evaluation indexes for each model obtained by the linear SVM for

- all words, and *2N* selected words with *N=1, 2, ..., 10, 20, ..., 70*, and
- the six types (*w.o, d.o, l.o, w.a, d.a* and *l.a*) of feature selection.

Figure 1 illustrates the relationship between the accuracy of each model obtained by the linear SVM and the value of *N*. The vertical axis represents for accuracy, and the horizontal axis is for the number of selected words *N*. The baseline represents the accuracy of the model by using all words.

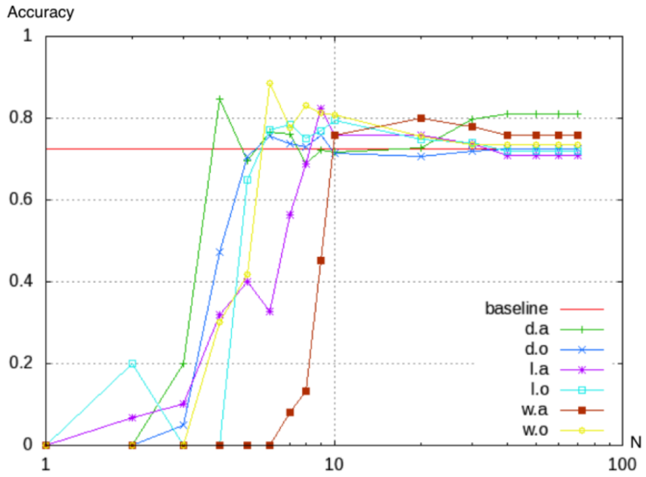

Figure 2. Accuracy

In the case of the model using all words, the accuracy was 0.7358. On the other hand, when *N=6*, applying feature selection *w.o*, the accuracy was 0.8853, which was the best score of all models. Thus, it seems that an appropriate feature selection based on linear SVM may be effective in reinforcing the estimation performance.

The scores for precision, recall, F-measure, and accuracy of the five models with the best scores are summarized in Table 3.

Table 3. The top five models and their precision, recall, F-measure, and accuracy scores.

Type of FS	N	precision	recall	F-measure	accuracy
w.o	6	0.7000	1.0000	0.7976	0.8853
d.a	4	0.7033	1.0000	0.8148	0.8470
w.o	8	0.7286	1.0000	0.8359	0.8300
l.a	9	0.6451	1.0000	0.7641	0.8224
w.o	9	0.6803	1.0000	0.8092	0.8139

Feature word

For the case of *N=6*, applying feature selection by *w.o*, we summarize the top six positive feature words and the top six negative feature words and their scores of importance in Table 4.

For example, the presence of the 12th week was the most influential learning activity in obtaining an A grade, and preparation and/or review of the 6th week was the most influential in failing to achieve an A grade.

We notice that *11b:o* is the second positive feature words, while *11b:x* appears as the third negative feature word. Thus, it can be suggested that preparation and/or review of the 11th week significantly distinguished A grade students from other students. In this "information science" course, the learning contents for the 11st week was bucket sort and binary search. It may therefore be supposed that

- these contents were the basis of other contents in the following weeks, or
- these contents were included in the final examination and were sufficiently important to classify the students' grades.

Focusing on negative feature words, the top three words were of the form *ib:x*, reflecting less than 600 seconds of slide browsing time of for preparation and/or review. The top three negative words shown that, in the middle part of the course, the students who neglected a preparation and/or review missed achieving an A grade. This result suggested that it may be important for a teacher to guide students in continuing to prepare for and review lectures through out the course, until the last week, in order to maximize their achievement.

These feature words may be used for two purposes. First, after finishing the course and grading the students, a teacher can tell the students which learning activities were not sufficient to obtain a good grade. Second, if a teacher is to conduct a similar course in the future, he/she can call students' attention to the learning activities indicated by positive feature words, and advise them to avoid learning activities indicated by negative feature words.

Table 4. The top six positive and negative feature words for *N=6*.

Positive		Negative	
Word	Score of word	Word	Score of word
12p:o	0.4554	*6b:x*	-1.0110
11b:o	0.4480	*8b:x*	-0.9019
10r:o	0.3223	*11b:x*	-0.8511
11r:x	0.2871	*8r:x*	-0.6693
5b:o	0.2686	*3t:x*	-0.6227
8p:x	0.2415	*13t:x*	-0.5108

Course grade scores vs. predicated A-score

For the case of *N=6*, applying feature selection by *w.o*, we let $f_1, f_2, ..., f_5$ be the classifiers of linear SVM, learned during the 5-fold cross validation. Then, we define the predicated A-score *pr(d)* of a document *d* of a student as the average of $f_1(d), f_2(d), ..., f_5(d)$.

For each student, we compared the grade score with the predicated A-score. The results are summarized in Figure 2, in which the vertical axis shows the predicated A-score, and the horizontal axis the grade score.

We can observe that there is a positive correlation between grade score and predicated A-score. Specifically, the correlation coefficient of them is 0.6333. Thus, it can be said that this model regarding whether or not students obtain an A grade is appropriate to discuss students' grade scores.

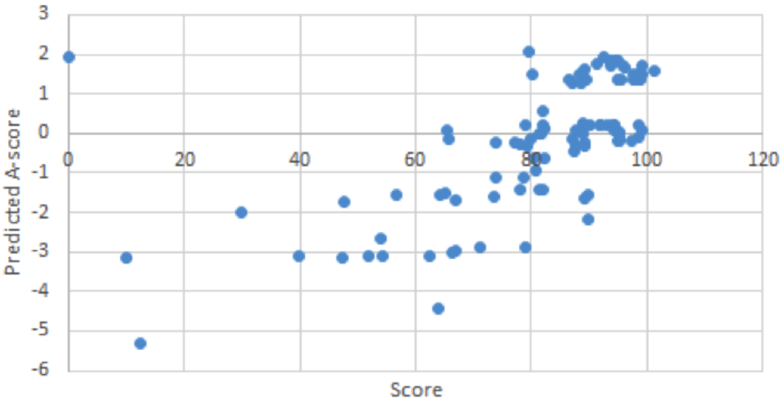

Figure 2. Course grade score vs. predicated A-score

Conclusion

In this paper, we proposed a method by which learning activities important for attaining high achievement in a course may be identified by using learning logs stored in an LMS and an e-book system. These logs contain the following four items, namely, attendance, slide browsing time for preparation and/or review, submission of reports, quiz scores, and grades. The learning activities of the students in a course can be represented by a vector that reflects achievement or non-achievement of each of the above four items in each week. In our method, first, linear SVM is applied to these vectors and a score of importance for the contribution of each learning activity to students' attainment of an A grade is calculated. Following this, we select N activities that have the best and worst scores of importance by following the six measures for evaluation. We then apply linear SVM to the data that consists of only the selected activities to verify that these activities are sufficient to infer a student's final achievement.

We applied this method to the data from 99 students attending an "Information Science" course. In the case of *N=6,* and applying feature selection *w.o,* the accuracy of prediction by using linear SVM was 0.8853, which was the best score of all constructed models. On the other hand, in the case of using of all learning activities, the accuracy was 0.7358. Hence, feature selection based on linear SVM appears to be effective in reinforcing the estimation performance. The selected activities were shown in section Table 3. From these results, we can observe that (i) preparation and/or review in the 11th week significantly distinguished A grade students from other students, and (ii) students who neglected preparation and/or review in the middle part of the course were unlikely to obtain an A grade. Furthermore, for each student, we compared the grade score with the predicated A-score by linear SVM with *N=6,* applying feature selection *w.o.* The correlation coefficient of them was 0.6333. Thus, it seems that the model regarding whether or not students obtain an A grade is appropriate in discussing the grade scores of students.

These results can be informative in telling students which learning activities were insufficient to obtain a good grade, and in advising students in following years of the same course on the important learning activities.

A number of issues remain to be investigated. Points of particular importance includes the following:

- More data from additional courses is required to support the present conclusions. It may be also interesting to compare the results of this study to data from another course.
- In our method, the thresholds for the achievement of slide browsing time for preparation and/or review and quiz scores were decided manually by the authors with no justification. It is important to determine the most suitable thresholds for identifying the specific features of the learning activities of high achieving students, automatically.
- By using our method, we can discover important learning activities for a good achievement. However, the reasons why these learning activities are selected by the model are not so easy to understand. Analysis of the relationships among learning activities, such as associations analysis, may help to interpret the present results further.

Endnotes

(1) http://booklooper.jp

(2) http://geta.nii.a.c.jp

(3) http://svmlight.joachims.org/

References

Baradwaj, B. & Pal, S. (2011) Mining Educational Data to Analyze Student's Performance, I*nternational Journal of Advanced Computer Science and Applications*, vol. 6, 2, pp. 63-69.

Cortes, C. & Vapnik, V. (1995) Support-Vector Networks, *Machine Learning*, vol.20, pp.273-297.

Dietterich, T.G., Lathrop, R.H. & Lozano-Perez, T. (1997) Solving the multiple instance problem with axis-parallel rectangles, *Artificial Intelligence*, Vol. 89, No.1-1, pp.31-71.

Goda, K., Hirokawa, S., & Mine, T. (2013) Correlation of grade prediction performance and validity of self-evaluation comments, *Proc. SIGITE'13*, pp. 35-42.

Hlosta, M., Herrmannová, D., Váchová, L., Kužílek, J., Zdrahal, Z. & Wolff, A. (2014) Modelling student online behaviour in a virtual learning environment, *Workshop Proc. LAK 2014,* 4 pages.

Ifenthaler, D. & Widanapathirana, C. (2014) Development and Validation of a Learning Analytics Framework: Two Case Studies Using Support Vector Machines, *Technology, Knowledge and Learning*, vol.19, pp.221-240.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K. & Yamada, M. (2015) E-Book-based Learning Analytics in University Education, *proc. ICCE2015*, pp.401-406.

Okubo, F., Shimada, A., Yin, C. & Ogata, H. (2015) Visualization and Prediction of Learning Activities by Using Discrete Graphs, *proc. ICCE2015*, pp.739-744.

Sakai, T. & Hirokawa, S. (2012) Feature Words that Classify Problem Sentence in Scientific Article, *Proc. iiWAS2012*, pp.360-367.

You, J. W. (2016) Identifying significant indicators using LMS data to predict course achievement in online learning, *Internet and Higher Education*, vol.29, pp.23-30.

Acknowledgments

The research results have been achieved by "Research and Development on Fundamental and Utilization Technologies for Social Big Data", the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan. The work of F. Okubo was in part supported by Kyushu University Interdisciplinary Programs in Education and Projects in Research Development No.27115.